

IEEE-SA Industry Connections White Paper

**IEEE 802 Nendica Report:
The Lossless Network for Data Centers**



IEEE | 3 Park Avenue | New York, NY 10016-5997 | USA

IEEE 802 Nendica Report: The Lossless Network for Data Centers

IEEE-SA Industry Connections



i

Copyright © 2018 IEEE. All rights reserved.



Trademarks and Disclaimers

IEEE believes the information in this publication is accurate as of its publication date; such information is subject to change without notice. IEEE is not responsible for any inadvertent errors.

Copyright © 2018 IEEE. All rights reserved.

IEEE owns the copyright to this Work in all forms of media. Copyright in the content retrieved, displayed or output from this Work is owned by IEEE and is protected by the copyright laws of the United States and by international treaties. IEEE reserves all rights not expressly granted.

IEEE is providing the Work to you at no charge. However, the Work is not to be considered within the “Public Domain,” as IEEE is, and at all times shall remain the sole copyright holder in the Work.

Except as allowed by the copyright laws of the United States of America or applicable international treaties, you may not further copy, prepare, and/or distribute copies of the Work, nor significant portions of the Work, in any form, without prior written permission from IEEE.

Requests for permission to reprint the Work, in whole or in part, or requests for a license to reproduce and/or distribute the Work, in any form, must be submitted via email to stds-ipr@ieee.org, or in writing to:

IEEE-SA Licensing and Contracts
445 Hoes Lane
Piscataway, NJ 08855

Comments on this report are welcomed by Nendica: the IEEE 802 “Network Enhancements for the Next Decade” Industry Connections Activity: <<https://1.ieee802.org/802-nendica>>

Comment submission instructions are available at: <<https://1.ieee802.org/802-nendica/nendica-lldcn>>

Comments will be considered for a future revision of this document and may stimulate the development of a new revision.

*The Institute of Electrical and Electronics Engineers, Inc.
3 Park Avenue, New York, NY 10016-5997, USA*

*Copyright © 2018 by The Institute of Electrical and Electronics Engineers, Inc.
All rights reserved. Published August 2018. Printed in the United States of America.*

IEEE and 802 are registered trademarks in the U. S. Patent & Trademark Office, owned by The Institute of Electrical and Electronics Engineers, Incorporated.

PDF: ISBN 978-1-5044-5102-4 STDVA23253

*IEEE prohibits discrimination, harassment, and bullying. For more information, visit
<http://www.ieee.org/web/aboutus/whatis/policies/p9-26.html>.*

No part of this publication may be reproduced in any form, in an electronic retrieval system, or otherwise, without the prior written permission of the publisher.

*To order IEEE Press Publications, call 1-800-678-IEEE.
Find IEEE standards and standards-related product listings at: <http://standards.ieee.org>*

Notice and Disclaimer of Liability Concerning the Use of IEEE-SA Industry Connections Documents

This IEEE Standards Association (“IEEE-SA”) Industry Connections publication (“Work”) is not a consensus standard document. Specifically, this document is NOT AN IEEE STANDARD. Information contained in this Work has been created by, or obtained from, sources believed to be reliable, and reviewed by members of the IEEE-SA Industry Connections activity that produced this Work. IEEE and the IEEE-SA Industry Connections activity members expressly disclaim all warranties (express, implied, and statutory) related to this Work, including, but not limited to, the warranties of: merchantability; fitness for a particular purpose; non-infringement; quality, accuracy, effectiveness, currency, or completeness of the Work or content within the Work. In addition, IEEE and the IEEE-SA Industry Connections activity members disclaim any and all conditions relating to: results; and workmanlike effort. This IEEE-SA Industry Connections document is supplied “AS IS” and “WITH ALL FAULTS.”

Although the IEEE-SA Industry Connections activity members who have created this Work believe that the information and guidance given in this Work serve as an enhancement to users, all persons must rely upon their own skill and judgment when making use of it. IN NO EVENT SHALL IEEE OR IEEE-SA INDUSTRY CONNECTIONS ACTIVITY MEMBERS BE LIABLE FOR ANY ERRORS OR OMISSIONS OR DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO: PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS WORK, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE AND REGARDLESS OF WHETHER SUCH DAMAGE WAS FORESEEABLE.

Further, information contained in this Work may be protected by intellectual property rights held by third parties or organizations, and the use of this information may require the user to negotiate with any such rights holders in order to legally acquire the rights to do so, and such rights holders may refuse to grant such rights. Attention is also called to the possibility that implementation of any or all of this Work may require use of subject matter covered by patent rights. By publication of this Work, no position is taken by the IEEE with respect to the existence or validity of any patent rights in connection therewith. The IEEE is not responsible for identifying patent rights for which a license may be required, or for conducting inquiries into the legal validity or scope of patents claims. Users are expressly advised that determination of the validity of any patent rights, and the risk of infringement of such rights, is entirely their own responsibility. No commitment to grant licenses under patent rights on a reasonable or non-discriminatory basis has been sought or received from any rights holder. The policies and procedures under which this document was created can be viewed at <http://standards.ieee.org/about/sasb/iccom/>.

This Work is published with the understanding that IEEE and the IEEE-SA Industry Connections activity members are supplying information through this Work, not attempting to render engineering or other professional services. If such services are required, the assistance of an appropriate professional should be sought. IEEE is not responsible for the statements and opinions advanced in this Work.

CONTENTS

ABSTRACT	1
EDITOR	1
CONTRIBUTORS/SUPPORTERS	1
OUR DIGITAL LIVES ARE DRIVING INNOVATION	2
TRENDS IN THE DATA CENTER	2
OnLine Data Intensive (OLDI) Services.....	2
Figure 1 – Parallel Application Hierarchy	3
Deep Learning.....	4
Figure 2 – Deep Learning Training.....	4
Figure 3 – Parallelism Tradeoff.....	5
NVMe over Fabrics.....	5
Cloudification of the Central Office	6
Figure 4 – Transition to Cloudified Architecture in the CO	7
Parallelism.....	7
Figure 5 – The Problem with Network Congestion	8
WHY TODAY’S DATA CENTERS AREN’T KEEPING UP	8
Figure 6 – Clos Network.....	9
Figure 7 – ECMP Load Balancing	9
Figure 8 – Current Congestion Management	10
ECMP collisions.....	11
Figure 9 – ECMP Load Balancing Collisions	11
ECN control loop delays	12
PFC head-of-line blocking.....	12
Figure 10 – PFC Head-of-Line Blocking.....	12
Lossless configuration complexity.....	13
Incast congestion	13
TECHNOLOGIES FOR THE FUTURE.....	14
Virtual Input Queuing	14
Figure 11 – Switch Packet Loss.....	14
Dynamic Virtual Lanes	15
Figure 12 – Dynamic Virtual Lanes	16

Load-Aware Packet Spraying.....	16
Figure 13 – Load Balancing Design Space.....	17
Push and Pull Hybrid Scheduling	17
Figure 14 – Push and Pull Hybrid Scheduling	18
STANDARDIZATION CONSIDERATIONS.....	19
CONCLUSIONS	20
CITATIONS	20

IEEE 802 Nendica Report: The Lossless Network for Data Centers

Abstract

Modern data centers are tasked with delivering intelligent multi-media responses to real-time human interactions. Massive amounts of data are being churned and sifted by highly parallel applications, such as Online Data Intensive Services (OLDI) and Artificial Intelligence (AI), which historically required specialized High-Performance Computing (HPC) infrastructure. New advancements in high-speed distributed solid-state storage, coupled with remote direct memory access (RDMA) and new networking technologies to better manage congestion, are allowing these parallel environments to run atop more generalized next generation cloud infrastructure. Generalized cloud infrastructure is also being deployed in the telecommunication operator's central office. The key to advancing cloud infrastructure to the next level is the elimination of loss in the network, not just packet loss, but throughput loss and latency loss. There simply should be no loss in the data center network. Congestion is the primary source of loss and in the network, congestion leads to dramatic performance degradation. This paper discusses the need for new technologies to combat loss in the data center network and introduces promising potential solutions.

Editor

Name	Affiliation
Paul Congdon	Tallac Networks

Nendica Chair

Name	Affiliation
Roger Marks	Huawei

Contributors/Supporters

Name	Affiliation
Jose Duato	Universitat Politecnica de Valencia
Barak Gafni	Mellanox
Feng Gao	Baidu
Liang Guo	CAICT
Jie Li	CAICT
Gu Rong	China Mobile

Contributors/Supporters continued on the next page

Contributors/Supporters continued

Richard Scheffenegger	NetApp
Mehmet Toy	Verizon
Sowmini Varadhan	Oracle
Jianglong Wang	China Telecom
Ilan Yerushalmi	Marvell
Yolanda Yu	Huawei

Our Digital Lives are Driving Innovation

For better or worse, our lives have been forever changed by digital technology. Digital technology is increasingly accessed and offered as a service from the cloud and, as a result, it is becoming more integrated with our natural lives.

Interacting with cloud services is now done in a human and natural way—through voice commands and visual recognition. In the not too distant future, as predicted by futurist Ray Kurzweil [1], the way we think will be augmented by the cloud. Already, services are personalized to our individual tastes by online data-intensive cloud services. We have come to expect instantaneous access to massive amounts of digital content by our very own voice commands. But how does all this work—in the backend—in the data center? How is it that huge quantities of data can be rendered into useful information within a timeframe that meets real-time human interaction delays?

The requirement to integrate digital technology into our daily lives is inspiring innovation in the data center. This innovation is driving the need for new levels of performance, scale, and reliability from the infrastructure. Enormous quantities of computing cycles are rendering massive amounts of data into real-time information and action. The delivery of information and action from the cloud data center needs to be fast! Consequently, the fabric within the data center needs to eliminate loss and deliver low latency and high throughput.

Trends in the Data Center

Application and storage architectures within the data center are continuously evolving to address increasing demands for real-time, interactive digital technology. Currently, four critical data center use cases are stressing the data center network. These include large-scale Online Data Intensive (OLDI) services such as automated recommendation systems for online shopping, social media and web searches; high-performance deep learning networks; modern telecommunication central office networks; and high-speed distributed pools of Non-Volatile Memory Express (NVMe) storage.

Online Data Intensive (OLDI) Services

The fundamental difference between Online Data Intensive services and their offline counterparts (such as MapReduce computations) is that they require immediate answers to requests that are coming in at a high rate. Latency control is a key concern. The end-user

experience is highly dependent upon the system responsiveness, and even moderate delays of less than a second can have a measurable impact on individual queries and their associated advertising revenue. A large chunk of unavoidable delay, due to the speed of light, is inherently built into a system that uses the remote cloud as the source of decision and information. This puts even more pressure on the deadlines within the data center itself. To address these latency concerns, OLDI services deploy individual requests across thousands of servers simultaneously. The responses from these servers are coordinated and aggregated to form the best recommendations or answers. Delays in obtaining these answers are compounded by delayed or 'straggler' communication flows between the servers. This creates a long tail latency distribution in the data center for highly parallel applications. To combat tail latency, servers are often arranged in a hierarchy, as shown in Figure 1, with strict deadlines given to each tier to produce an answer. If valuable data arrives late because of latency in the network, the data is simply discarded, and a sub-optimal answer may be returned. Studies have shown that the network becomes a significant component of overall data center latency when congestion occurs in the network [2].

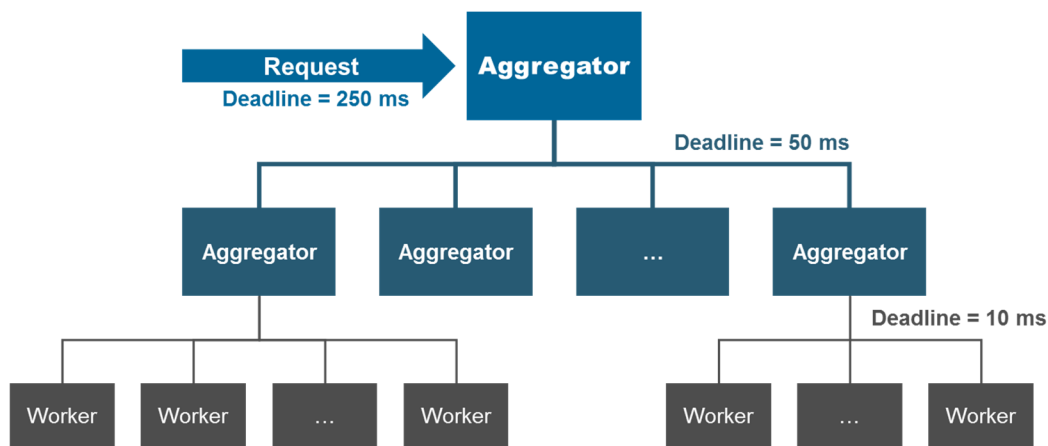


Figure 1 – Parallel Application Hierarchy

The long tail of latency distribution in OLDI data centers can be caused by various factors [3]. One is simply related to the mix of traffic between control messages (mice) and data messages (elephants). While most of the flows in the data center are mice, most of the bytes transferred across the network are due to elephants. Therefore, a small number of elephant flows can delay the set-up of control channels established by mice flows. Since OLDI data centers are processing requests over thousands of servers simultaneously, the mix and interplay of mice and elephant flows is highly uncoordinated. An additional complexity is that flows can change behavior over time; what was once an elephant can transform into a mouse after an application has reached steady state. Another cause of latency is due to incast at the lower tiers of the node hierarchy. Leaf worker nodes return their answers to a common parent in the tree at nearly the same time. This can cause buffer over-runs and packet loss within an individual switch. It may invoke congestion management schemes such as flow-control or congestion notification, which have little effect on mice flows and tail latency.